RESEARCH FUNDING

Sustaining the Data and Bioresource Commons

Globalization of biomedical research requires sustained investment for databases and biorepositories.

Paul N. Schofield,[1,2]* Janan Eppig,[2] Eva Huala,[3] Martin Hrabe de Angelis,[4] Mark Harvey,[5] Duncan Davidson,[6] Tom Weaver,[7] Steve Brown,[8] Damian Smedley,[9] Nadia Rosenthal,[10] Klaus Schughart,[11] Vassilis Aidinis,[12] Glauco Tocchini-Valentini,[13] John M. Hancock[8]

Development of powerful, high-throughput technologies, together with globalization of scientific research, presents the biomedical research community with unprecedented challenges for the management, archiving, and distribution of data and bioresources (*1*). We need a social contract between funding agencies and the scientific community to accommodate "bottom-up" integration and "top-down" financing of databases and biorepositories on an international scale.

The biological commons is evolving away from a traditional differentiated structure to one in which origination, ownership, and distribution of data and materials are subsumed by the same community (*2*). Scientific progress depends on efficient and open sharing to generate maximum value (*3*–*5*). The traditional paradigm of sharing scientific data and results through the published literature is no longer effective where new technologies produce large volumes of diverse types of data and biological materials. Critical to the maintenance, distribution, and archiving of these data and materials, therefore, are stable public databases and repositories. Provision of public funding for these long-term repositories does not fall into the traditional model of science funding, yet financial support is vital if we are to maximally exploit the investment into experimental science. Although funding agencies may exhort their experimental investigators to develop a "dissemination plan" for the data and bioresources they develop, in reality, such requirements are often not fulfilled, and noncompliance has little or no consequence. This often means that funders are effectively washing their hands of responsibility for future accessibility and reuse of the data and bioresources whose generation they have financed. Instead, funding for data and bioresource repositories needs to be ring-fenced from hypothesis-driven research and supported sufficiently to ensure preservation and maintenance of its outputs.

The contents of the new generation of data and bioresources are continuously being enhanced and augmented by the community of user-producers. There is not a sequential phase of research, followed by storage and use. Databases need continually to revise their data models to accommodate new data types. The associated bioontologies and other informatics tools need to continue to be developed, maintained, and applied to data to standardize and maximize access, retrieval, and exploitation for discovery. Repositories also need to innovate and respond to emerging disruptive technologies. Consequently, any distinction between time-delimited research projects and long-term, relatively static infrastructures is being eroded. The traditional distinction between "infrastructure" and "research" is even less appropriate, presenting a challenge to those funders who continue to think in these terms. The additional value created by manual data curation and integration in databases like Mouse Genome Informatics (MGI) or the *Arabidopsis* Information Resource (TAIR) is enormous, yet this activity does not fall into the recognized domain of "research activity" for many agencies.

The scale of investment required across the life sciences may be estimated from current funding of large community databases and bioresources. For 2009, MGI's core activity funding was U.S.$6.3 million including overhead, and for TAIR was $1.6 million. Curation activities alone of EMAGE, the embryonic mouse in situ gene expression database (*6*) based in the United Kingdom, currently cost roughly $0.7 million per year. The range represented here reflects the scope, as well as amount and complexity, of data.

The amount of investment in databases worldwide is a disproportionately small fraction of overall research budgets. For example, as little as a 5% allocation of the U.S. National Institutes of Health spending of $20.9 billion that constituted research grant awards in 2009 (*7*) would provide $1 billion toward biological data resources and huge added value for the community. Although some large public databases have survived rounds of competitive renewal, others have failed, often as a result of funding policy decisions rather than poor projects. Among high-profile databases, the one with the most recent funding crisis

[1]Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, CB2 3EG, UK. [2]The Jackson Laboratory, Bar Harbor, ME 04609, USA. [3]Carnegie Institution for Science, Stanford, CA 94305, USA. [4]Institute of Experimental Genetics, Helmholtz Zentrum München, D-85764, Munich, Germany. [5]Department of Sociology, University of Essex, Colchester, CO4 3SQ, UK. [6]Medical Research Council (MRC), Human Genetics Unit, Western General Hospital, Edinburgh, EH4 2XU, UK. [7]MRC Mary Lyon Centre, Harwell, Didcot, Oxfordshire, OX11 0RD, UK. [8]MRC Harwell, Mammalian Genetics Unit, Oxfordshire, OX11 0RD, UK. [9]European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK. [10]European Molecular Biology Laboratory, Monterotondo, I-00015, Rome, Italy. [11]Department of Infection Genetics, Helmholtz Centre for Infection Research, and University of Veterinary Medicine, Braunschweig, D-38124, Germany. [12]Biomedical Sciences Research Center Alexander Fleming, GR-16672, Vari, Athens, Greece. [13]Istituto di Biologia Cellulare, CNR, I-00015 Monterotondo Scalo (Rome), Italy.

*Author for correspondence. E-mail: ps@mole.bio.cam.ac.uk

is the *Arabidopsis* TAIR database (*8*), with other important databases, such as Eurexpress or Euregene, currently facing crises at the end of their existing funding cycles. Pressure will continue to grow as new community projects come to fruition and new data types (such as magnetic resonance images) without a dedicated public archive, need to be in the public domain.

Most biological stock centers have begun charging for products and services to meet their costs, but still require subsidy for long-term sustainability. For example, the annual operating costs of the European Mutant Mouse Archive network, including the repository and associated database (*9*), are €5 million (~U.S.$7 million), of which €2 million (U.S.$ 2.8 million) are provided by the European Commission. U.S. repositories receive NIH subsidies; for example, in 2009 the Knockout Mouse Project (KOMP) repository received $3.4 million and the Jackson Laboratory Mouse Mutant Resource received $1.5 million (*7*).

Various models for financial and scientific sustainability have been tried and discussed. "Recover costs from users" is a frequent exhortation from funding agencies. Experience shows that the viability of such strategies for databases is illusory (*10*). Public databases depend on the community's contributing freely to the commons, the quid pro quo being open and free access. No example of a successful fee-for-service model organism database exists.

In a recent online survey conducted by TAIR (*11*), users were strongly against the possibility of having to pay for access. Exclusion of investigators from access to data and resources, disadvantaging those most unable to pay (especially investigators in developing economies), was the primary reason cited. Second, data integration, increasingly an essential aspect of data sharing, would be crippled by the inability to integrate data between databases which did and did not charge for access. The seamless network of data (*12*) would be fragmented and disabled.

We have already seen bottom-up initiatives for data standardization and sharing on a global scale and acknowledgement of the importance of the commons (*3*–*5*, *13*). The challenges for funding agencies are those of coordination and strategy: how to adequately recognize the transnational nature of data and bioresources in funding instruments and how to sustainably fund core international resources when funding sources are likely to remain predominantly national.

An example of the advantages of international cooperation can be seen in the International Knockout Mouse Consortium (IKMC) (*14*), which is generating knockouts of protein-coding genes and distributing vectors, embryonic stem cells, and knockout mice. It involves centers in Europe, the United States, and Canada. International coordination has promoted efficiency and more rapid delivery of resources to investigators and has avoided duplication. Another is the International Nucleotide Sequence Database Collaboration, an action supported by the National Center for Biotechnology Information (NIH), European Molecular Biology Laboratory, and the Japanese National Institute of Genetics.

Funded nationally, required internationally, whose responsibility is it to fund future access to data and resources produced by projects such as the IKMC at the end of project funding? This generic problem is recognized by the European Council and Commission, which established the European Strategy Forum on Research Infrastructures (ESFRI) program (*15*, *16*), designed to identify strategic research road maps needed across the sciences and to build international organizations to coordinate and receive national funding. One of these projects, Infrafrontier (*17*), has as its goal the large-scale systematic phenotyping and archiving of mouse models to support not only the European, but also the international, biomedical research community. The success of this innovative program depends on the willingness of national agencies to support novel transnational organizations. Negotiations between the partners are under way, and Canada has recently joined the Infrafrontier project as the first full non-European partner.

We propose that national funding agencies should initiate infrastructure coordination programs, analogous to the European ESFRI process, from which support of internationally important databases and repositories might be sought. These funding opportunities should be restricted to data and bioresource-sharing infrastructures. The programs would implement the shared national research priorities of the agencies and would reflect strengths or needs in particular fields. This model already has fledgling examples. For example, in plant biology, the International Steering Committee on Plant Genomics (ISCPG) (*18*) includes representatives from funding agencies in Australia, Brazil, Canada, China, Japan, European Union, Consultative Group on International Agricultural Research, United States, and United Kingdom. The mission statement of the ISCPG could be taken as a model for international infrastructure coordination activity.

In this proposal, representatives of the national funding agencies and the scientific community would consult on research infrastructure priorities and needs in a particular area, and agencies whose policy priorities match those needs would cooperate on shared international funding calls. A stumbling block to universal participation in this model is that many national agencies are currently legally unable to provide funds to researchers in other countries. The adoption of a legal framework providing an international legal identity, such as the recently developed European Research Infrastructure Consortium (ERIC) (*19*) for internationally integrated projects, would solve this and could facilitate the mobilization of national funds.

International harmonization of data sharing and intellectual property policy could be both necessary and highly advantageous. Intellectual property right considerations can be addressed through negotiation and agreements between individual national funding organizations; these are now common and do not present insuperable problems. Interagency cooperation, for example on material transfer agreements, may accelerate the adoption of common policies to the great advantage of the scientific community.

References and Notes

1. F. S. Collins, *Science* **327**, 36 (2010).
2. M. Harvey, A. McMeekin, *Public or Private Economies of Knowledge* (Edward Elgar, Cheltenham, 2007).
3. Creative Commons, sciencecommons.org.
4. Toronto International Data Release Workshop Authors, *Nature* **461**, 168 (2009).
5. P. N. Schofield *et al.*, *Nature* **461**, 171 (2009).
6. EMAGE, www.emouseatlas.org/emage.
7. Research Portfolio Online Reporting Tools, report.nih.gov/.
8. A. Abbott, *Nature* **462**, 258 (2009).
9. P. Wilkinson *et al.*, *Nucleic Acids Res.* **38** (Database issue), D570 (2010).
10. C. Chandras *et al.*, *Database* **2009**, bap017 (2009) (Oxford).
11. Survey results, arabidopsis.org/doc/about/tair_survey/411.
12. R. W. Williams, *Front. Neurosci.* **3**, 154 (2009).
13. D. Field *et al.*, *Science* **326**, 234 (2009).
14. International Mouse Knockout Consortium, *Cell* **128**, 9 (2007).
15. I. W. Mattaj, *Nature* **465**, 1005 (2010).
16. ESFRI, ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri.
17. Infrafrontier, www.infrafrontier.eu/.
18. ISCPG, www.iscpg.com/index.htm.
19. Directorate-General for Research, *Legal Framework for a European Research Infrastructure Consortium (ERIC): Practical Guidelines* (European Commission, Brussels, 2010); ec.europa.eu/research/infrastructures/pdf/eric_en.pdf.
20. This Policy Forum is based on discussions at a meeting of the Coordination and Sustainability of International Mouse Informatics Resources (CASIMIR) consortium held 11 and 12 November 2009 in Rome and supported by the European Commission, contract no. LSHG-CT-2006-037811. A list of participants is available at *Science* Online.

Supporting Online Material

www.sciencemag.org/cgi/content/full/330/6004/592/DC1

10.1126/science.1191506